SAMPLE DESIGN IN THE FINNISH AGRICULTURAL INCOME STATISTICS

Paavo Väisänen

SAMPLE DESIGN IN THE FINNISH AGRICULTURAL INCOME STATISTICS

Paavo Väisänen
P.O.Box 3A, FIN-00022 Statistics Finland

Abstract

The Finnish Agricultural Income Statistics, published yearly by Statistics Finland, are based on a survey in which the data are collected from the farms in connection with taxation. The sampling design is stratified simple random sampling, in which Neyman allocation is used to calculate the sample sizes for the strata. The Farm Register is used as the sampling frame where variables such as region, production sector and arable land are available for stratification. The total incomes of farms from the previous survey serve as the allocation variable. Stratification and Neyman allocation rendered the estimates of most income variables more effective when measured by the design effect (DEFF) values which ranged from 0.3 to 0.7. Ratio estimation was studied by using arable land as an auxiliary variable. The sample was also evaluated by calculating estimates for variables available from administrative records and by comparing them with the true values. The estimated values were systematically bigger than the true values. Non-response among small farms was one reason for this systematic error. A comparison by production sector revealed that the biggest differences were in cattle farming and in the production of cereals. An examination of the correlations in these sectors revealed a linear dependence between the survey and auxiliary variables. Ratio estimation was used in these sectors to reduce the error of estimates and to balance the variables known from other sources.

KEY WORDS: Income of agriculture, multipurpose survey, ratio estimation, stratified sampling

1. Introduction

In Finland the same sample is used to survey both production and income in agriculture. The Information Center of the Ministry of Agriculture and Forestry (ICMAF) publishes agricultural production statistics and Statistics Finland maintains incomes statistics on agriculture. Both these statistics are published yearly and they are based on surveys. The distributions of the variables describing agricultural production and income are skew, and a part of the skewness is taken into account by stratifying according to production sector and arable land. In multipurpose surveys a number of different aspects have to be considered and a design that satisfies simultaneously as many research needs as possible should be chosen. In most cases it is difficult to find suitable variables for efficient sample designs, such as sampling with probabilities proportional to size.

The sampling design is based on stratified simple random sampling (STRSRS). The Farm Register was used as the sample frame. It comprises all farms and is updated annually. The data on the income statistics of agriculture were gathered from tax returns and by using questionnaires appended to tax returns.

In the estimation, the sampling design was evaluated by comparing true and estimated values calculated using different methods. The suitability of the ratio estimator was tested by using the Farm Register of 1992 and 1993. The sample was drawn from the 1992 register, but since results concerned the year 1993, the ratio estimators were studied using the 1993 register. π-expanded estimates were calculated to arable land, and they were compared to the true totals, calculated by production sector from the register. When compared to the π-expanded estimates, the totals from the 1993 farm register were for most variables bigger than the true values. Correlation coefficients were calculated for the arable land and the study variables by production sectors and size classes. Cattle farming and cereal production had the highest dependencies and the biggest differences, and ratio estimation was used in the strata connected to these production sectors.

The population changes yearly, which has been taken into account in the estimation. Constant coefficients were used to adjust the population numbers of the sampling time to correspond to the survey time. The non-response model is based on the assumption that the distributions are the same in the non-response group as in the respondents group. In this naive model the estimation is based on the response group in each stratum (Särndal et al., 1992).

In surveys, precision of estimates depends essentially on the sampling design and on the use of auxiliary information in the estimation. Different sampling designs and estimation strategies were compared when searching for methods and variables to render the estimation design more efficient. Ratio estimators were useful for variables with linear dependence to the arable land under cultivation. The structure of linear dependence was studied by calculating the regression equations and testing intercepts. Point plots were drawn, suggesting the idea of skewed distributions. Random and total errors were evaluated by comparing the point estimates calculated in the sample and response groups. Non-response errors were studied by using data from the Farm Register. Non-response was found to have skewed distributions related mainly to small farms.

The survey strategy consists of a rotating panel, in which one third is changed each year except in the strata of large farms, some of them stay in the sample every year and some rotates slower than in the main sample. A new rotation group is sampled every year which is the same kind of sample of population as the preceding rotation groups. In 1992 the sample design was renewed, and the stratification structure was changed. Lower limits were set for the population of holdings. The new design was introduced as one rotation group at a time, and the sample of 1993 included one rotation group, sampled in 1991 by using the earlier sample design.

2 Sampling design

The population of income statistics of agriculture is the sub-group of the population of ICMAF survey. In the income statistics the population comprises farms having arable land under cultivation two hectares or more except in the hay and cereal production sectors

where the lower limit for the arable land was three hectares. The population of the ICMAF survey consists of farms which have arable land at least one hectare. Institutional farms were excluded from the both surveys.

The Farm Register rendered possible to stratify according to production sector, arable land, livestock and different regional classifications as rural districts. The variable of a arable land was used to measure the size of the farm and the hay and cereals production, and in this production sector farms were classified into five sub-classes according to size. The big holdings, in Finland ones with over 100 hectares of arable land under cultivation, were classified into three strata according to the major districts. Small production sectors and some rural districts were combined together to form larger strata, and the final number of strata reduced to 155 (Statistics Finland, 1995).

The sample was allocated according to Neyman allocation. The income of farms was used as the allocation variable, which was commensurable variable in all production sectors. The allocation calculations were based on data from the previous survey. Allocation according to income was not the best possible in the case of the cereal production or animal husbandry, where distributions were skewed concentrating only on the strata of certain production sectors (Väisänen, 1993).

The sample size was 14 627 farms. Tax forms were returned by 93 % of holdings, and 72 % of those returned the statistical questionnaires. The sampling frame comprised 119 055 holdings belonging to the population. From the point in time when the sample was selected, in 1992, to the point in time of the survey, in 1993, there were changes in the population. For instance, in 1993 there were 4316 holdings less than in 1992, making a total of 114 739 holdings for the population size. In the sample this caused an overcoverage of 703 farms. Some holdings had finished production, some had been linked to other holdings, owner changes took place, there were changes in the area of arable land and forest area, just to mention a few reasons. If these changes were not to be taken into account, this would cause the systematic bias in the results. The sample was matched to the Farm Register of 1993, where 77 farms were found that had not been included in the register. This was due to undercoverage of the sampling frame. The annual changes in the population cause over- and undercoverage in the sample, which cannot be avoided, but the bias can be adjusted by weighting using coefficients depending on the ratios of the population sizes in successive years.

The population of the ICMAF survey comprised 4 575 more holdings than in the survey of Statistics Finland. The difference was due to the fact that the ICMAF also included holdings with areas of arable land of only one to two hectares in the sample in some production sectors. The difference in sample sizes was 479 holdings.

3. Estimation

Sampling design was taken into account in the estimation procedure. All holdings were

provided with appropriate weights, which were used in the estimation. The weights included factors of sampling probability, non-response adjustment and changes in the frame. In addition to this, in the production sectors of cattle farming and cereal production, the ratio estimators were used which resulted in an additional factor in the weights.

Let y denote a survey variable and x an auxiliary variable received from the register, and y_i the value of y for the unit i, and x_i, respectively. Let π_{hi} be the probability that a unit i belongs to the sample in stratum h, and N_h the size of population in stratum h and n_h the sample size. Now $N=\Sigma N_h$ and $n=\Sigma n_h$. The π-estimator of the total in stratified sampling is (Särndal et al.,1992)

$$\hat{t}_{ySTR} = \sum_{h=1}^{H} \frac{y_{hi}}{\pi_{hi}} \qquad (1)$$

where the number of strata H=155. In STRSRS design $\pi_{hi}=\pi_h=n_h/N_h$ in all strata.

In the annual changes in the population were adjusted by constants $k_{h,t} = N_h' / N_{h,t-1}$ which are the ratios of the populations in the strata in points of time t and t-1. A common method for nonresponse adjustment is weighting in homogeneous groups. In this case strata were selected into homogeneous groups. Response probabilities P_{rh} were calculated in the sample $P_{rh} = v_h/n_h$ where v_h is the number of the respondents. When entering these factors to expression (1) we got for the estimator of total t_{ySTR} (Särndal et al.,1992)

$$\hat{t}_{ySTR} = \sum_{h=1}^{H} k_h^{93} \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi} P_{rh}} = \sum_{h=1}^{H} N_h^{93} \sum_{h=1}^{n_h} \frac{y_{hi}}{v_h} \qquad (2)$$

Standard errors were calculated for the most significant variables. The estimator of the variance in the STRSRS design is (Särndal et al., 1992)

$$var(\hat{t}_{ySTRSRS}) = \sum_{h=1}^{H} \frac{N_h^2(1-f_h)}{n_h} \left[\frac{\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2}{n_h-1}\right] \qquad (3)$$

where for simplifying notations $N_h = N_{h,t}$ and $n_h = v_h$.

Standard errors were calculated using the SUDAAN software (Shah et al.,1991). SUDAAN uses the general Taylor linearization method in the estimation of standard errors of means in subgroups.

The sample design was evaluated by using the data with which in the first stage the sample expansion weights or π-weights were calculated (Särndal et al.,1991), and the ratio estimators were studied both in the whole data and in the subgroups.

The estimators and the sample was evaluated by using variables available form the Farm Register. Results in Table 1 indicates a systematic error in the data. A part of this overestimation was caused by non-response bias (section 5) that the non-response weighting did not correct. When corresponding characteristics were calculated by the production sectors the errors were discovered to vary in size and direction. The values were too big in cattle farming and cereal production. The values estimated were 65 000 hectares bigger for dairy production, 11 700 for beef production, 13 400 for cereal production and 6 500 hectares smaller in other crop production.

Table 1 Comparison of the estimated and true values for the register variables (in hectares)

Variable	Register	Estimate	Error
Cultivated arable land	2 210 807	2 308 736	97 929
Rented arable land	398 423	427 778	29 355
Forest land	4 707 460	4 904 737	197 277

4. The use of variables from the Farm Register in adjusting the survey data

All study variables that were linearly dependent on arable land, were estimated more efficiently by using arable land as an auxiliary variable. At the same time the estimate of the total of arable land was balanced to respond the true value. Ratio estimators were used in the production sectors where linear dependence occurred between the variables studied and the arable land, and π-estimators were used in the strata where no dependence was observed. The ratio estimator corrected the error of the estimate of arable land to 30 000 hectares.

Correlation coefficients of the most significant variables are shown in Appendix 1. Correlations were in general rather small varying from 0.1 to 0.2 in most income variables, except in cereals, where correlation was 0.6. Correlations in the variables that involved crops were rather high. In the production sectors of dairy and beef production, the correlations of income variables were 0.5 and 0.4. In the production sectors of cereal and hay production the correlation on crop income was 0.8. Correlations were calculated using SAS software.

The point plots of survey variables, based on the area of arable land, formed flabellate patterns concentrating near the origin (Appendix 3). In the production sectors the point plots appeared to be linear in the groups which had positive values for study variables.

The examination of correlation coefficients and graphs gave support to each other. The ratio estimator was not useful in the whole sample but only in production sectors in which differences between estimated and true values were big. Further demands were the dependence between study and proxy variables. The use of ratio estimation balanced the register variables nearer the true values and the variables studied were estimated more efficiently. A minor degree of error remained in the estimates of the register variables. The correlations and graphics did not support the total balancing of register variables which might have caused systematic errors to variables studied.

A separate ratio estimator of the total t_{yR} was used in the strata of cattle farming and crop production. The auxiliary variable x was the area of arable land. The ratio estimator in stratum h is expressed as (Särndal etal.,1992)

$$\hat{t}_{yR_h} = \hat{R}_h t_{hx} = \frac{\hat{t}_{hy}}{\hat{t}_{hx}} t_{hx} \qquad (4)$$

The estimator of variance of the estimator (4) is calculated using Taylor linearization and the approximate variance of the ratio estimator of the total is (Särndal et al.,1992)

$$\begin{aligned} AV(\hat{t}_{hyR}) &= \left(\frac{t_{hx}}{\hat{t}_{hx}}\right)^2 \sum\sum\left(1-\frac{\pi_k\pi_l}{\pi_{kl}}\right)\left(\frac{y_{hk}-\hat{R}_h x_{hk}}{\pi_k}\right)\left(\frac{y_{hl}-\hat{R}_h x_{hl}}{\pi_{hl}}\right) \\ &= \left(\frac{t_{hx}}{\hat{t}_{hx}}\right)^2 \frac{1-f_h}{n_h} \sum\sum\frac{(y_{hk}-\hat{R}_h x_{hk})(y_{hl}-\hat{R}_h x_{hl})}{n-1} \end{aligned} \qquad (5)$$

where $f_h = n_h/N_h$.

Let H' be the group of the strata where π-estimator was used and H'' the strata where ratio estimation was used. The estimator of total t_y is now

$$\hat{t}_y = \sum_{h\in H'} \hat{t}_h + \sum_{h\in H''} \hat{R}_h \hat{t}_h \qquad (6)$$

The estimator of the variance for this estimator is calculated respectively with expression (3) when $h \in H'$ and with (5) when $h \in H''$

5. Non-response

The non-response rate was 7.4 % (1035 holdings). The reasons for non-response were, for instance, that tax returns were not received. The tax returns were not received for holdings where the owner had an other holding in some other municipality of permanent

residence and where the tax returns were made. Register data are available also for the non-response group so it was possible to study the distribution of non-response.

Table 2 The means (in hectares) of arable land in response and non-response groups

Reason for non-response	Number of holdings	Mean of arable land
Responses	12 889	21.24
Belongs to other municipality	204	12.95
Form was not returned	82	17.23
Additional form missing	174	11.94
Received only basic forms	10	11.89
Generation changed	90	27.54
Owner changed	63	25.10
Bankrupt	8	12.95
Owner was not reached	23	19.11
Inadequate tax return	38	31.03
Form missing	343	16.70

Small holdings were more likely to belong into the non-response group than large holdings (Table 2). So the responses largely represented bigger holdings, which was one reason why the estimates of arable land exceeded the true values.

The non-response rate for the statistical questionnaire was 29 %. Data from the tax returns and the Farm Register were used for imputation of missing data in the statistical questionnaires. The expenditure variables of farms were imputed by grouping responses into homogeneous subgroups and missing expense values were placed with estimated values. Items of expenditure and income were estimated as percentages of the totals received from tax returns. Homogeneous groups were formed according to production sectors and farm size classification (Statistics Finland, 1995).

6. Efficiency of the sample design

Standard errors, coefficients of variation (CV) and design effects (deff) were calculated for the estimates (Appendix 2). In multipurpose surveys the efficiency of the sampling design varies depending on the subject under study. The design effect is defined as a ratio of the design based variance and the variance of simple random sample. Deff numbers can be used to compare how well the sample design functions for different variables. The deff numbers varied for income variables from 0.3 to 1.5 and for the items of expenditure from 0.4 to 1.0. CV varies from 1 % to 3 % for most variables. Exceptionally high CV values

were obtained for poultry (4,2 %), other livestock production (8,9 %), potatoes (4,8) and sugar beet (6,9 %), which all have skewed distributions in regional sense, as the sample includes only few units which have this kind of production.

7. Conclusions and future work

The sample design is considered as an example of probability sampling in multipurpose surveys. The precision of the estimates satisfies the needs of users in general. The sample included one rotation group in which the old sample design was used and the effects of this were not studied. In the sample of 1994, all rotation groups are sampled using the method presented in this paper. Data collection on the income statistics of 1994 was carried out in the autumn of 1995, and we have plans to combine the data of the production survey and the income survey to get a larger picture of the sample design. Because Finland joined the European Union this year, several new subsides are available for farmers who are obliged to give more detailed information about their production and the use of arable land. All these data are added to the Farm Register, which will offer new possibilities for the use of register data in sampling and estimation.

References

Statistics Finland (1995). The Business and Income Statistics of the Farm Economy 1993. Agriculture and Forestry 1995:1. Helsinki (in Finnish, English summary).

Information Center of Ministry of Agriculture and Forestry (1994). Farm register 1992. *Agriculture and Forestry* 1994:1. Helsinki (in Finnish)

National Board of Agriculture (1993). *Monthly Review of Agricultural Statistics,* 1993: 6 and 10. Helsinki: Government Printing Centre

SAS Institute Inc.(1989). SAS language and Procedures: Usage, Version 6, First Edition, Cary, NC

Shah B.V., Barnwell B.G.,Hunt P.N, LaVange L.M. (1991). *SUDAAN User's Manual, Release 5.50.* Research Triangle Park, NC,27709

Särndal C.E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling.* Springer-Verlag

Väisänen P. (1993). The sample design of Finnish agricultural surveys. *1993 ICES proceedings of the Section on Establishment Surveys,* American Statistical Association.

Appendix 1 The correlation coefficient for some income and expenditure variables with arable land

Variables	Whole sample	Cattle farming	Cereals product.
Income from sale of agricultural products	0,6	0,7	0,7
Livestock production	0,4	0,7	0,2
- Dairy products	0,3	0,5	0,0
- Beef	0,2	0,4	0,1
- Pork	0,2	0,2	0,1
- Poultry	0,1	0,0	0,1
- Other livestock products	0,0	0,1	0,1
Crop production	0,6	0,5	0,8
- Cereals	0,6	0,5	0,8
- Potatoes	0,1	0,1	0,1
- Sugar-beet	0,2	0,2	0,2
- Garden products	0,0	0,0	0,1
- Other crops	0,2	0,2	0,3
Other income	0,4	0,5	0,6
- Subsides	0,5	0,5	0,7
- Subsidy based on area	0,0	0,0	-0,0
- Subsidy based on field area	0,9	0,8	0,9
- Other subsidies	0,3	0,3	0,5
Total income	0,7	0,8	0,8
- Reserves credit to income	0,2	0,0	0,2
- Other agricultural income	0,2	0,3	0,2
Wages	0,4	0,4	0,4
Purchase of input	0,4	0,4	0,5
- Livestock	0,2	0,3	0,1
- Feed, etc	0,3	0,4	0,2
- Other livestock production costs	0,3	0,4	0,1
- Fertilizers and lime	0,6	0,6	0,7
- Seed, herbicides, pesticides, etc	0,5	0,4	0,6
- Tools and equipment for short-term use	0,3	0,4	0,4
Other expenditure	0,7	0,8	0,7
- Fuel and lubricants	0,5	0,6	0,6
- Electricity	0,5	0,5	0,5
- Machinery repair	0,5	0,6	0,6
- Maintenance of ditches, etc	0,3	0,3	0,4
- Building repair	0,3	0,3	0,3
- Land rents	0,4	0,4	0,5
- Other rents	0,2	0,2	0,2
- Other deductible expenditure	0,5	0,6	0,6
Depreciation	0,6	0,6	0,6
Total expenditure	0,6	0,7	0,7
Gain or loss from agriculture	0,6	0,4	0,5
Assets of farm economy	0,6	0,6	0,6
Liabilities of farm economy	0,5	0,5	0,5

Calculated by SAS

Appendix 2 Taxable income and expenditure, and assets and liabilities, in agriculture in 1993, FIM million; standard error and deff

		SE	CV %	Deff
Income from sale of agricultural products	21505,5	175,3	0,8	0,43
Livestock production	15637,5	156,8	1,0	0,42
- Dairy products	7249,8	97,6	1,3	0,68
- Beef	3491,3	64,5	1,8	0,77
- Pork	3591,3	90,6	2,5	0,27
- Poultry	1146,4	48,3	4,2	0,29
- Other livestock products	158,7	14,2	8,9	0,42
Crop production	5868,0	66,5	1,1	0,33
- Cereals	3894,2	42,3	1,1	0,29
- Potatoes	468,3	22,6	4,8	0,43
- Sugar-beet	482,4	33,6	6,9	1,07
- Garden products	467,5	26,1	5,6	0,39
- Other crops	555,6	21,8	3,9	0,46
Other income	4799,6	52,8	1,1	1,04
- Supplementary non-agricultural activity	389,3	25,0	6,4	1,16
- Subsides	2940,5	26,6	0,9	1,01
- Subsidy based on area	656,7	6,2	0,9	0,93
- Subsidy based on field area	962,3	7,4	1,8	0,54
- Other subsidies	1321,5	22,8	1,7	1,25
- Reserves credit to income	311,4	8,8	2,8	0,91
- Other agricultural income	1158,4	29,3	2,5	1,04
Total income	26305,1	192,3	0,7	0,46
Wages	529,9	16,6	3,1	0,40
Purchase of input	8731,5	96,2	1,1	0,39
- Livestock	1403,4	41,4	2,9	0,41
- Feed, etc	3547,1	47,8	1,3	0,38
- Other livestock production costs	629,5	9,0	1,4	0,51
- Fertilizers and lime	2027,4	18,0	0,9	0,55
- Seed, herbicides, pesticides, etc	925,0	10,4	1,1	0,43
- Tools and equipment for short-term use	199,1	3,1	1,5	0,73
Other expenditure	5700,9	44,5	0,8	0,51
- Fuel and lubricants	713,7	8,6	1,2	0,67
- Electricity	462,9	4,4	0,9	0,68
- Machinery repair	898,2	10,3	1,1	0,69
- Maintenance of ditches, etc	140,4	2,4	1,7	0,62
- Building repair	244,9	5,4	2,2	0,88
- Land rents	255,2	5,9	2,3	0,31
- Other rents	283,3	5,9	2,1	0,60
- Other deductible expenditure	2701,9	23,8	0,9	0,48
Depreciation	2603,8	25,5	1,0	0,74
- Buildings	678,0	8,7	1,3	0,73
Equalization reserve	568,8	11,1	1,9	0,88
Total expenditure	18135,0	148,3	0,8	0,42
Gain or loss from agriculture	6418,4	64,9	1,0	0,74
Capital income	1173,6	16,7	1,4	1,03
Earned income	5244,8	56,8	1,1	0,70
Assets of farm economy	19901,2	181,2	0,9	0,74
Liabilities of farm economy	19898,1	253,1	1,3	0,68

Statistics Finland, 1995. Calculated by SUDAAN

Appendix 3 The point plots of income from cattle farming and cereal and hay production according to cultivated arable land

Figure 1 Dairy production

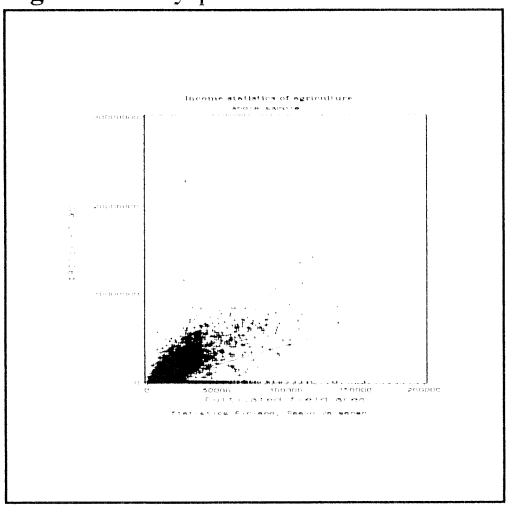

Figure 2 Beef production

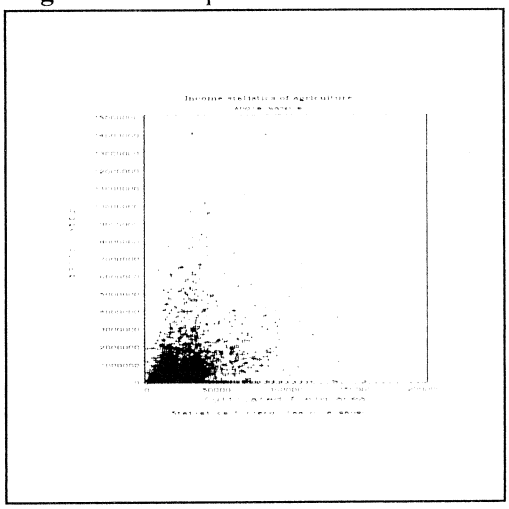

Figure 3 Dairy production in the subgroup of cattle farming

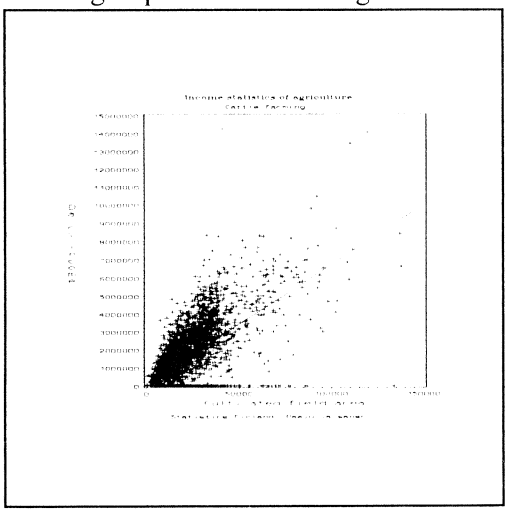

Figure 4 Beef production in the subgroup of cattle farm.

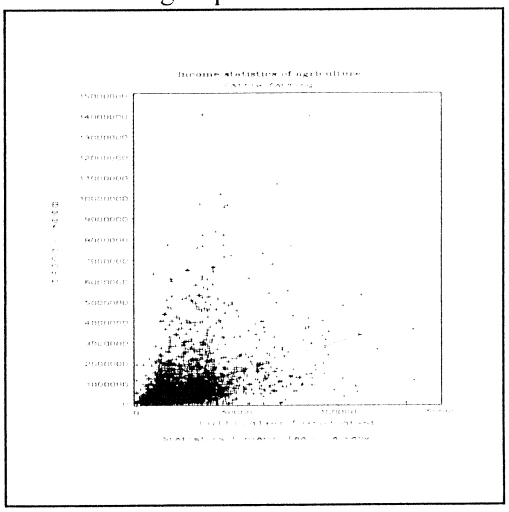

Figure 5 Income from creal productio,

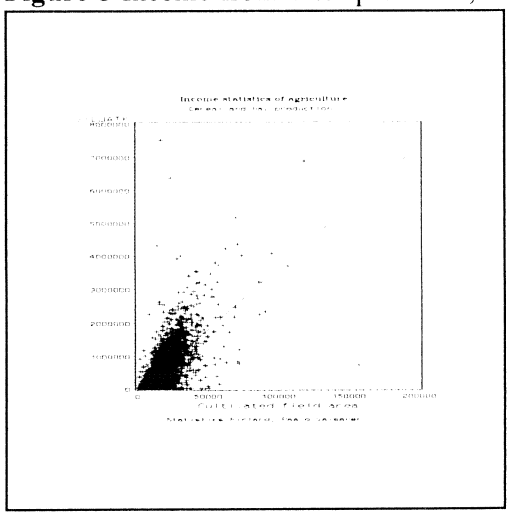

Figure 6 Income from sale of agr.